CA Journal of Cultural Analytics

The Tell-Tale Hat: Surfacing the Uncertainty in Folklore Classification

Peter M. Broadwell, David Mimno and Timothy R. Tangherlini

02.08.17

Peer-Reviewed By: Chad Wellmon

Clusters: Genre

Journal ISSN: 2371-4549

Article DOI: 10.22148/16.012

Dataverse DOI: 10.7910/DVN/MUYD7T

Classification is a vexing problem in folkloristics. Although broad genre classifications such as "ballad", "folktale", "legend", "proverb", and "riddle" are well established and widely accepted, these formal classifications are coarse and do little more than provide a first level sort on materials for collections that can easily include tens, if not hundreds, of thousands of records.[1] Many large collections of folklore have been classified using systems designed for very specific tasks, usually related to early theories about the spread of folk narrative.[2] Beyond the straightforward parsing of folk expressions into easily recognized formal genres (e.g. ballad, riddle, joke, legend, fairytale, etc.), the overarching emphasis of these schemas is on topic indexing. Perhaps best known of these indices is the ATU index of fairy tales, designed to assist scholars who are interested in the comparison of fairy tales from one or more cultures.[3] Another well-known index, the motif index of folk literature compiled by Stith Thompson (1955-58), is designed to assist scholars in discovering the relationships between complete narratives and their component parts, as well as the movement of motifs across time and space, where the motif is conceptualized as "the smallest element in a tale having a power to persist in tradition."[4] Other genre specific classification schemes include the Migratory Legend [ML] catalog[5] and *Danmarks gamle Folkeviser* [DgF].[6] More collection specific indices include the Child Ballads[7] and, of particular interest to this study, the typological indices to the Danish folklore collector Evald Tang Kristensen's legend collections.[8]

Nearly all these classification schema assign a single label to a story.[9] If a story can fit into multiple categories, the

[1] For example, see John Zemke, "Editor's Column. Archives, Databases, and Special Collections," Oral Tradition 28, no. 2 (2014): 1-3; Desislava Paneva, Konstantin Rangochev, and Detelin Luchev, "Ontological Model of the Knowledge in Folklore Digital Library." In Proceedings of the Fifth HUBUSKA Open Workshop on Knowledge Technologies and Applications, ed. Tatiana Urbanova, István Simonics, and Radoslav Pavlov (Kosice, Slovakia: HUBUSKA, 2007), 47–55.

[2] Kaarle Krohn, Die folkloristische Arbeitsmethode begründet von Julius Krohn und weitergeführt von Nordischen Forschern, Instituttet for sammenlignende kulturforskning Ser. B 5 (Oslo: H. Aschehoug & co, 1900–1902); Antti Aarne, Verzeichnis der Märchentypen, FF Communications 3 (Helsinki: Suomalainen Tiedeakatemia); Alan Dundes, "From Etic to Emic Units in the Structural Study of Folktales," Journal of American Folklore 75, no. 296 (1962), 95.

[3] Hans-Jörg Uther, The Types of International Folktales: A Classification and Bibliography, Based on the System of Antti Aarne and Stith Thompson, FF Communications 284, 3 vols. (Helsinki: Suomalainen Tiedeakatemia, 2004).

[4] Stith Thompson, The Folktale (New York: Holt, Rinehart, Winston, 1946), 415-416.

[5] Reidar Christiansen, The Migratory Legends, FF Communications 175 (Helsinki: Suomalainen Tiedeakatemia, 1958).

[6] Svend Grundtvig, Axel Olrik, Hakon Grüner-Nielsen, Karl-Ivar Hildeman, Erik Dal, Iørn Piø, Thorkild Knudsen, Svend Nielsen, and Nils Schiørring, eds., Danmarks gamle Folkeviser, 12 vols. (Copenhagen: Universitets-jubilæets danske Samfund, 1966–1976; orig. 1853–1976).

[7] Francis James Child and George Lyman Kittredge, The English and Scottish Popular Ballads (Boston: Houghton Mifflin, 1882-1894).

[8] Evald Tang Kristensen, Danske sagn, som de har lydt i folkemunde, udelukkende efter utrykte kilder, 7 vols. (Aarhus: Aarhus Folkeblads Bogtrykkeri, 1892–1901; repr. 1980); Evald Tang Kristensen, Gamle folks fortællinger om det jyske almueliv, som det er blevet ført i mands minde, samt enkelte oplysende sidestykker fra øerne, 6 vols. (Kolding: Sjødt and Weiss, 1891-1894).

[9] James Abello, Peter M. Broadwell, and Timothy R. Tangherlini, "Computational Folkloristics," Communications of the Association for Computing Machinery 55, no. 7 (2012): 60-70. In the case of the motif index, stories are assigned a chain of motifs. But each motif is only assigned a single number

CA Journal of Cultural Analytics

person performing the classification makes a judgment call, assigning to the story a label from a predetermined list that he or she considers to be the best fit. Indeed, apocryphal stories of the origins of the motif index tell of Stith Thompson "sit[ting] at his work table in Room 40 of the old Indiana University library, piles of motif slips laid out on the table before him, and spin[ning] a poker chip container around in lazy Susan fashion until he found the right slot for the slip under consideration. But if it seemed an isolated and questionable motif he discarded it, with brisk and decisive judgment."[10] In this regard, Thompson was little different than the scholars before him, from Aristotle to Linnaeus, who had struggled with the uncertainty inherent in placing things into categories. His classification work also echoes that of the Danish collector, Evald Tang Kristensen, many decades earlier, who, writing in the preface to his second printed collection of legends, said, "I must ask the reader for forbearance on several fronts. First, the ordering of the stories, which has its difficulties; but it should be noted that where a story is obtuse or distorted, then the classification is based on a judgment. For example tale 388 (classified in "On prophecy and portents") could have been put in section ix ("Fairytale-like legends"), 389 ("On prophecy and portents") in section iv ("On revenants and all types of ghosts"), and 438 and 439 ("Religious legends") in section vi ("On witchcraft")."[11]

Most folklore indices have significant gaps and unusual lacunae. One need only consider Thompson's classification practices that led to the motif index to understand why. Thompson worked with little hesitation, discarding motifs he considered *sui generis*, eliding the grey areas between categories and discarding potentially interesting outliers. In many classification systems, as Tang Kristensen himself notes, the labels are somewhat arbitrary. In addition, because they are applied in a manual fashion by people whose attention can wander, who can get tired, and who can find the work tedious and underappreciated, the application of the labels is inconsistent. Consequently, it is often difficult to search comprehensively in most collections: resources are hard to locate, and there is no guarantee that the located resources constitute a comprehensive retrieval of the relevant materials housed in the archive.

We suggest that computational approaches to the indexing of folklore collections might offer some relief to the artifacts of past practices. Although the ATU, ML and similar indices all have their time honored uses, and will continue to be useful for the specific purposes for which they were designed, more flexible indexing systems can assist in resource discovery and can support research questions not considered by the original indexing regimes. Indeed, in modern folklore research, most research questions fall outside the bounds of the narrowly conceived classification schema devised by earlier scholars. While we do not propose to discard the important and useful work that has gone into classifying folklore collections, the computational methods considered below can significantly augment and extend these earlier classifications, while opening up additional means for discovering resources in these complex collections. At the same time, our use of computational tools differs from standard commercial or scientific use. Our goal is not to treat existing classifications as "ground truth" labels and build machine learning tools to mimic them, but rather to use computation to better quantify the variability and uncertainty of those classifications.

A well known concern in folklore is the tension between *etic* and *emic* categories of classification, perhaps best articulated by Alan Dundes.[12] In his consideration of Kenneth Pike's[13] largely linguistic distinction between etic and emic units, Dundes points out that the motif index, the ATU and, by extension, nearly all topic-based classifiers rely on highly variable aspects of content (etic units) as opposed to more stable morphological or structural features (emic units). Consequently, "classification is not based upon the structure of the tales themselves so much as the subjective evaluation of the classifier... If a tale involves a stupid ogre and a magic object, it is truly an arbitrary decision whether the tale is placed under II A, Tales of Magic (Magic Objects), or II D, Tales of the Stupid Ogre."[14] He continues, "Perhaps the best illustration of the fact that Aarne-Thompson typology is based upon the variable and not upon the constant may be found by examining tale types which differ only with respect to the dramatis personae. In the Animal Tale (Type 9), The Unjust Partner, there is a version listed in which in the division of the crop, the fox takes the corn while the benighted bear takes the more bulky chaff. Under the Tales of the Stupid Ogre, one finds Tale Type 1030, The Crop Division. It is the same story except that the dramatis personae are a man and an ogre."[15] Adding emic or structural considerations to a classifier would allow one,

(in other words, a motif can only be "about" one thing). We use the word "story" loosely to refer to any folkloric utterance.

[10] Richard Dorson, "Stith Thompson (1885-1976)," Journal of American Folklore 90, no. 355 (1977): 2-7.

[11] Evald Tang Kristensen, Sagn fra Jylland, samlede af folkemunde, Jyske folkeminder især fra Hammerum herred 4 (Copenhagen: Karl Schønbergs forlag, 1880), i. Abbreviated as JFm 4.

[12] Alan Dundes, "From Etic to Emic Units in the Structural Study of Folktales," Journal of American Folklore 75, no 296(1962): 95–105.

[13] Kenneth L. Pike, Language in Relation to a Unified Theory of the Structure of Human Behavior, Part I (Glendale: Summer Institute of Linguistics: 1954).

[14] Dundes, "From Etic to Emic Units," 98.

[15] Dundes, "From Etic to Emic Units," 98.

CA Journal of Cultural Analytics

for instance, to capture what Dundes has called the "motifemic" equivalence across tales or tale types, so that the fox/man and bear/ogre distinctions would not be the discriminating factors for classification as in the ATU index.[16]

Ethnographic studies tell us that storytellers in a target culture internalize emic categories, which are part of the rich context of lived experience, and, consequently, their storytelling can exhibit great genre and topical (read etic) variation. Intrinsically, storytellers recognize that stories are about many different things at once, a polysemy that most classification schema, particularly those based on etic features, fail to capture. Similarly, storytellers' conceptual categories and the boundaries defining those categories, influenced as they are by the dynamics of tradition and social change, are often fluid. It is in this porousness between boundaries that a great deal of cultural negotiation occurs and, consequently, it is in these border regions that one may find the most interesting stories. Hard to classify stories or motifs--those that live in the borderlands, so to speak-- arise when storytellers push at the conservative tendencies of their tradition groups.[17] As such, these stories may help capture moments of cultural change and should not be discarded "with brisk and decisive judgment".

In 1878, relatively early in his collecting and publishing career, Tang Kristensen wrote to Svend Grundtvig, his mentor and the leading folklorist of the day, that "no classification scheme can substitute for a comprehensive indexing, as one finds more often than not in a single legend more than one thing to consider", thereby indicating his growing uneasiness with the necessary yet unfortunate classificatory regime under which he had placed the stories of his informants. Ultimately, these classic methods of Linnaean-inspired systems fail, given their lack of flexibility and, as Dundes points out, their reliance on largely etic features. Tang Kristensen, in his postscript to *Danske sagn*, recognizes as much, writing, "In regards to the organization of the material, I have but little to say. It is more or less the same system as I have followed since the start, because I find it to be the most practical, particularly when you have a great deal of material with which to work. There are probably better systems that could have been devised for scientific use, but the one used here can't be completely objectionable..."[18] Recognizing that some type of classification is better than none, and, in amusing echo of George Box's recognition that "essentially, all models are wrong, but some are useful", Tang Kristensen makes short work of selecting his approach over those of earlier scholars including the Dane, Just Matthias Thiele, and the Norwegian, Andreas Faye, whose classification scheme he considered to be "fundamentally as unfortunate a system as one could imagine."[19]

The goal of our work, of course, is not to do away with Tang Kristensen's existing classifiers, but to augment his approach by finding those instances where the boundaries between categories are permeable, thus capturing those moments of etic judgment which inevitably elide the emic complexity of storytelling. These moments might tell us a great deal about the impact of social, economic and political change on the conceptual categories of tradition participants. Below, we address this concern by considering the "one story-one label" problem, which we propose to solve using a ranked list of possible labels for any given story. Importantly, we discover that stories that do not fit easily into a single category, or stories that span multiple categories, are particularly interesting. Taking a cue from van Gennep's concept of liminality, we call these borderline cases "liminal" stories, and propose that the very features that make them hard to categorize make them interesting to researchers.[20]

Target corpus

In the following work, we focus on folk narratives collected by the Danish folklorist, Evald Tang Kristensen, from 1867 to 1924.[21] Over the course of his six decade long career, Tang Kristensen collected roughly 160,000 stories, songs, riddles, jokes, recipes and other descriptions of daily life from nearly four thousand individuals in western Denmark. His complete published collection consists of eighty-four volumes that follow a highly idiosyncratic and inconsistent classification scheme. Our study corpus represents a small fraction of his overall collection, consisting of approximately 31,000 legends and descriptions of everyday life.[22] We focused on this subset of the corpus as it is the least "noisy" in regards to OCR

[16] Alan Dundes, The Morphology of North American Indian Folktales, FF Communications 195 (Helsinki: Suomalainen Tiedeakatemia 1964; repr. 1980).

[17] Walter Anderson: Kaiser und Abt. Die Geschichte eines Schwanks, FF Communications 42 (Helsinki: Suomalainen Tiedeakatemia, 1923).

[18] Tang Kristensen, Danske sagn, vol. 7, 493.

[19] Tang Kristensen, Danske sagn, vol. 7, 493.

[20] Arnold van Gennep, Les rites de passage (Paris: Emile Nourry, 1909).

[21] Timothy R. Tangherlini, Danish Folktales, Legends, and Other Stories (Seattle: Univ. Washington Press; Copenhagen: Museum Tusculanum, 2013).

[22] Timothy R. Tangherlini, "'It happened not too far from here…': A Survey of Legend Theory and Characterization," Western Folklore 49(1990): 371-390.

CA Journal of Cultural Analytics

errors, and has the most complete metadata related to storytellers and places. These stories were largely published in his two main collections of legends[23] and descriptions of daily life among the Jutlandic peasantry.[24] In broad terms, Tang Kristensen arranged his collection first by genre and then by topic. Topics were further broken into sub-categories. For example, in one published collection, *Danske sagn*, which is dedicated to the distinct genre of legends, the first-level topic of "hidden folk" has several second-level classifications, including "hidden folk one has either seen or heard".

Figure 1. Printed volumes in the Tang Kristensen collection and their physical labels.

After transcribing his field notes into fair copy, Tang Kristensen would arrange the stories according to his classification schema and then, once he felt the groupings were sufficiently large (a calculation based on printing costs and volume size), send the fair copy to be printed. These collections were published on an ongoing basis as he collected more and more material. This unusual work flow has led to a secondary problem for people who want to use the collection: someone interested in legends about ghosts and hauntings, for instance, needs to consult not only *Danske Sagn* [Danish legends], a seven volume collection, but also the six volumes of *Danske sagn, ny række* [Danish legends, new series], as well as various volumes in *Jyske folkeminder* [Jutlandic folklore]. Even then, there is no guarantee of comprehensiveness, as stories about ghosts and hauntings also appear in the twelve volumes of *Jyske Almueliv* [Jutlandic Peasant Life], other smaller collections not arranged in larger multi-volume series, and the many unpublished, and thus unindexed, stories that round out the collection.

Summarizing his work on *Danske Sagn* and *Jyske Almueliv* and how he made the decision to put stories in one collection or the other, Tang Kristensen wrote, "When this work is finished, and one wants to compare the two works with each other, not only will one discover many of the same informants in the two collections, but one will frequently find points of connection in the content that are at times so close to one another that something found in one collection could just as easily have been placed in the other. In particular, this concerns parts of the fourth volume of the legend collection and parts of the second and fifth volumes of the work on descriptions of everyday peasant life. But there is an important difference, and that is

[23] Tang Kristensen, Danske sagn; Evald Tang Kristensen, Danske sagn, som de har lydt i folkemunde, samlede og for størstedelen optegnede af Evald Tang Kristensen, ny række, 7 vols. (Copenhagen: Woels Forlag, 1928–1939).

[24] Tang Kristensen, Gamle folks fortællinger om det jyske almueliv; Evald Tang Kristensen, Gamle folks fortællinger om det jyske almueliv, som det er blevet ført i mands minde, samt enkelte oplysende sidestykker fra øerne. Tillægsbind, 6 vols. (Aarhus: Forfatterens forlag, 1900–1902).

CA Journal of Cultural Analytics

really noticeable in those sections of the legends that contain the old folk beliefs. The legends should really focus in general on belief, and the descriptions of peasant life on life as it is lived."[25] Since the stories can only appear under a single rubric, and are only ever published once (if at all), and since there is no cross-indexing of the published volumes in the collection, many stories that include mention of themes such as ghosts and hauntings are not easily discoverable. Consequently, our work also shows how one can concatenate indexing across multiple, unconnected collections, in addition to demonstrating how one can identify "liminal" stories.

Reverse engineering a folklore expert

Ultimately, the goal of our work is to reverse engineer the folklore expert by replaying the moment of classification. In the case of Tang Kristensen, he had unparalleled domain expertise, but was also constrained by time and the necessity of publishing, as the sale of published volumes supported subsequent collecting work. As a result, the moment of classification was one that mediated between the exigencies of folklore collecting and publishing in late nineteenth century Denmark, the need to align the collection with the expectations of the scholarly community, represented largely by Svend Grundtvig and his efforts to classify the Danish ballads, and the expressive culture of the thousands of people from whom Tang Kristensen collected. This classificatory regime led to several failures, such as top level categories that were at times overly broad (e.g. "Life outdoors"), and second level categories that were at times overly precise (e.g. "Funeral processions one has seen, or that pass one by" and "Funeral processions one has met or followed"). At the same time, the classification regime necessarily cut across the native, emic classification of stories, erasing storytellers' conceptions of stories and worldviews, focusing entirely on surface level, etic considerations. Consequently, there are several distorting factors in the classification scheme itself: the emic classifications that are elided through topic (etic) classification, and the imprecision of top level topic classifications or the overly precise nature of second level classifications.

In writing the description of our reverse engineering work below, we deliberately avoid terms that are commonly used in Machine Learning, where labels are "true," "correct," or "gold standard." This linguistic distinction highlights the fundamentally different perspective that humanists have on classification as a tool. Our goal is not to create a system that mimics the decisions of a human annotator, but rather to better represent the porous boundaries between labels and identify the piles on which a story *could* have been placed over a century ago late on a cold wintry night in a dimly lit schoolhouse in eastern Jutland. We note the contrast between our use of computers to problematize existing distinctions and the common concern in the Humanities that computers deal only with binaries and black-and-white distinctions.

Prior work on classification in Danish folklore

In prior work, we considered classification problems in a small subsection of this corpus. Using a subcorpus comprising 948 stories, we focused on creating a flexible system for navigating through the collection making use of (a) the existing classification systems, referred to as the ETK indices, (b) a shallow, two-level ontology devised by Tangherlini[26] and applied manually to the corpus, and (c) a feature set for each story derived from a simple "bag-of-words" model of the texts.[27] We showed that a hypergraph model of the stories offered researchers new opportunities for discovering stories that otherwise were difficult to discover using existing classification schema. Similarly, we showed that seemingly unrelated phenomena could share surprising features.[28] The visually rich navigation interface also allowed for a sophisticated navigation through the complex "story space" of the collection. Predicated on the notion of "homophily"—that similar things like to gather in the same neighborhood—the story space navigator allows a user to explore groupings of "similar" stories. We used standard similarity calculations (cosine similarity; Hellinger distance) calculated on a weighted feature set to create these neighborhoods of similar stories. We incorporated aspects of this classification system into the Danish Folklore Nexus[29]

[25] Tang Kristensen, Danske sagn, vol. 7, 493.
[26] Timothy R. Tangherlini, Interpreting Legend: Danish Storytellers and Their Repertoires (New York: Garland Publishing, 1994; repr. 2015).
[27] Abello et al., "Computational Folkloristics."
[28] Abello et al., "Computational Folkloristics," 68-69.
[29] See figure 2, and Timothy R. Tangherlini and Peter M. Broadwell "Sites of (re) Collection: Creating the Danish Folklore Nexus," Journal of Folklore Research 51, no. 2 (2014): 223-247.

CA Journal of Cultural Analytics

corpus exploration interface, affording users of the Nexus the opportunity to discover related stories that might not share the same topic indices.

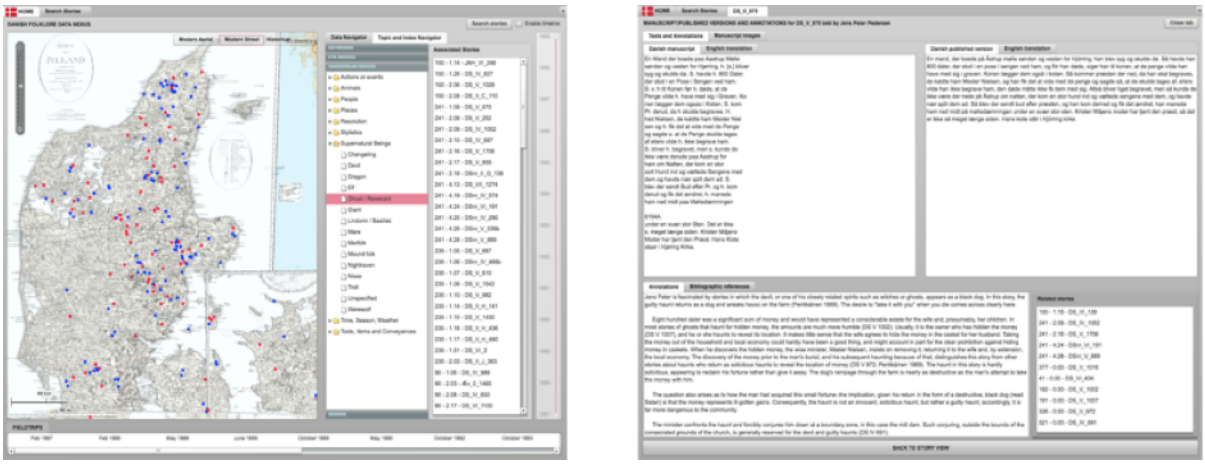

Figure 2. The Danish Folklore Nexus interface (www.purl.org/danishfolktales) highlighting the corpus exploration interface (left) and the machine-learning classifier/recommender in the related-stories window (right).

The current work is intended to extend these discussions of automated folklore classification to a larger selection of stories. Below, we explain how we generate a multi-level, ranked list of possible classifications for each story, information that can be incorporated in the visual navigation of the story space or in future iterations of the Danish Folklore Nexus.[30] We recognize that machine learning systems are generally "dumb", and consequently we develop an approach that expands the set of possible classifications suggested by the actions of the original expert.

Problem and methods

We present here a core problem related to the issue of classification of a large folklore corpus. The "One Story-One Label" problem can be stated as follows: Given a pre-established range of possible labels (here, those devised by Tang Kristensen), is the existing label the best one? We extend this simple question, and consider whether one can devise a ranked list of possible labels for the story, thereby recognizing that, even given a relatively small range of possible labels, a story is rarely only about one thing. For this problem, we devised a ranked labeling system that makes use of a probabilistic classifier. A feature of this system is that it is agnostic to genre. Unlike most folklore classifiers, which do not allow for the mixing of genres, this approach allows a ballad about a murder to share a label or labels with a legend about that same murder.

For this problem, we are particularly interested in (a) stories that straddle numerous categories (a problem of which Tang Kristensen was acutely aware) and (b) words that are highly associated with multiple labels. These borderline cases suggest a degree of ambiguity that likely allows the story to be used to address aspects of ambiguity in daily life in nineteenth century Denmark, thus capturing the emic categories that inform folk belief and expression. We build a dynamic visualization of the classification space (see figure 4), based on the notion of the "confusion matrix", that allows a user to rapidly discover the stories that span multiple classifications. This interface also presents users with the top-ranked alternative classification and highlights the terms that are most strongly indicative of the original classification versus those that suggest the alternative category.[31]

[30] Discussed in Abello et al., "Computational Folkloristics."

[31] The interface can be accessed at http://bit.ly/2jIdu6o. Access to the second level indices are available at two additional sites: http://bit.ly/2ct3y9q and http://bit.ly/2ckqhaE

CA Journal of Cultural Analytics

Data preparation

Most folklore collections require significant preprocessing to be machine-readable. The eighty-four volumes of Tang Kristensen's published collection were scanned and OCR'ed using ABBYY FineReader software tuned to the printing conventions of the books through training, and to the orthographic conventions of late nineteenth century Danish through a custom dictionary. Individual stories were separated into individual text files, and the resulting stories and their attendant metadata (collection, volume, storyteller, place collected) were stored in a MySQL relational database. A selection of 948 stories was translated into English, while the remaining ~30,000 stories were stored solely in Danish. An overview of the collection can be found in *Danish Folktales, Legends and Other Stories*.[32] Numerous infelicities still exist in the corpus, although these tend to be relatively consistent (mistaking "u" for "n"), related to shifting orthographic conventions across the published volumes (ø vs ö, å vs aa), or related to Tang Kristensen's attempts to capture dialect.

A story and some context

To get a sense of these stories, it is necessary to accompany Tang Kristensen on one of his many forays into the countryside to collect stories. On one such field trip in 1888, Tang Kristensen encountered Jens Korregård, a farm owner and parish bailiff, from the small northern village of Havbro, who, along with his brother Niels, told numerous stories to Tang Kristensen, including the following:

> Per Overlade was out one evening shooting hares. It was up on Kræn Møller's field. Kræn was in the process of moving his farm, and the old farm had not been completely disassembled yet, and Per intended to hide amid the old frame that was still standing and shoot a hare or two. But when he gets there, he sees an old man who is sitting in there with a red cap on who nods to him. Per gets scared and doesn't dare go in there, and so he doesn't catch any hares.[33]

Tang Kristensen classified the story with the label "Hidden Folk", and placed it in the sub-category of "Hidden Folk one has either seen or heard". It is not entirely clear why Tang Kristensen chose the label he did for the story. It is possible that he intended to include the story in his collection of stories about Jutlandic peasant life but failed to get the story copied fair prior to the compilation of that work, where it would likely have been categorized as a story about "life outdoors." The only indication that the old man may be one of the hidden folk is Per's reaction to the sudden appearance of the man, as opposed to anything the old man does. Although one might be tempted to imagine that Tang Kristensen was eager to pad the volume of hidden folk, since this volume was the first in his collection and by far the largest, it seems just as likely that he assigned the story to the category of "hidden folk" since the category was essentially a catch-all classification for uncanny stories about encounters in liminal areas, particularly fields.[34]

For any story, one might ask, which features would a Danish farmer likely have noticed in this story? Borrowing from de Certeau's notion that stories represent "repertoires of schemas of action" and the recognition that storytellers use stories to comment on and explore the implications of changes in their social and physical environments, the features of the story to which an individual would attend were likely influenced by this ongoing narrative negotiation of positionality vis-à-vis these factors.[35]

During the nineteenth century, rural Denmark underwent a profound transformation. After the dissolution of the manorial system at the end of the eighteenth century, the increasing quest for farming efficiencies led many large landholders to support efforts to sell off their underperforming fields to their former tenants while maintaining their lucrative forest-based hunting and logging rights.[36] As part of the shift to smaller, private farms, formerly centralized villages were dismantled, and the farm buildings were moved out onto the newly partitioned fields. This reapportionment of the fields and the

[32] Tangherlini, Danish Folktales, Legends, and Other Stories.

[33] Tang Kristensen, Danske sagn, vol. 1, 14.

[34] Peter M. Broadwell and Timothy R. Tangherlini, "GhostScope: Conceptual Mapping of Supernatural Phenomena in a Large Folklore Corpus," in Maths meets Myths, ed. Ralph Kenna, Máirín MacCarron, Padraíg MacCarron (Cham, Switzerland: Springer, 2017), 131-158.

[35] Michel de Certeau, The Practice of Everyday Life, trans. Steven Rendall (Berkeley: The University of California Press, 1984), 23.

[36] Timothy R. Tangherlini, "Annotation to DS_V_202", in Danish Folktales, Legends and Other Stories: The Danish Folklore Nexus, Digital Materials (Seattle: University of Washington Press; Copenhagen: Museum Tusculanum, 2013); Tangherlini, "Annotation to DS_VII_285", in Danish Folklore Nexus.

CA Journal of Cultural Analytics

subsequent removal of buildings lasted for many decades, tapering off in the waning decades of the nineteenth century.[37] The result was nothing less than a radical spatial reorganization of the Danish landscape.

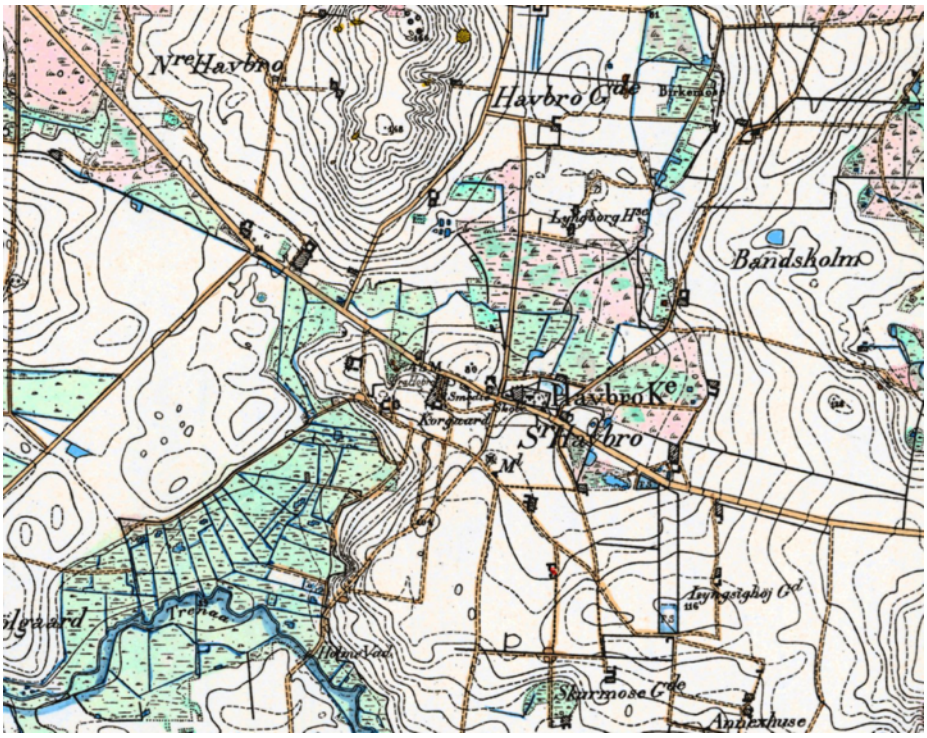

Figure 3. The fields around Havbro, where Jens Korregård lived and Kræn Møller had his original farm.

References to these changes often emerged, albeit obliquely, in the traditional storytelling of Danish farmers, the very people caught up in these currents.[38] As Michael Taussig notes in his study of Bolivian copper miners, people, when confronted with radical change in the local environment, often appeal to traditional storytelling to help understand the implications of these changes.[39] This appeal to the folkloric also obtained in the Danish situation, and gave rise in the 1800s to countless stories commenting on the shifting terrain of economic organization.

Seen in this context, Jens Korregård's story likely addresses many other subjects than Tang Kristensen's classification suggests, and it may not necessarily have as its main focus hidden folk. Indeed, the potentially supernatural figure of the old man sitting among the timbers of the old farm appears to have far more in common with the *nisse*, the Danish household spirit, than hidden folk, and likely belongs, at least in the first instance, among stories about household spirits.[40] The old man's red cap is particularly telling, as it is most often associated with household spirits in Danish tradition. Interestingly, the story makes oblique reference to poaching, a liminal activity that clearly sets the stage for a supernatural encounter.[41] To wit, Per Overlade is hunting on someone else's fields and, to a Danish nineteenth century farmer, the reference is anything but subtle. Furthermore, the disassembled farm points directly to the period of the reapportionment of the fields, a situation that was rife with economic and social pitfalls. It is little wonder, then, that a *nisse*, whose purpose is to protect the integrity of the farm, would appear to scare off the would-be theft of the rabbit (while a person was not allowed to hunt in the largely private forests, animals that crossed their land were "fair game"), and protect the farm in its vulnerable, disassembled state. At the same time, the deliberate disassembly of the farm building may have confused the *nisse*--as well as the storytellers--and presaged a shift in storytelling concerning the previously stable category of the household spirit with his removal to the fields that were once the exclusive domain of the hidden folk. Given the potentially vicious nature of the household spirit,

[37] Tangherlini, Danish Folktales, Legends, and Other Stories, 11.

[38] Timothy R. Tangherlini, "The Beggar, the Minister, the Farmer, his Wife and the Teacher: Legend and Legislative Reform in Nineteenth Century Denmark," in Legends and Landscape, ed. Terry Gunnell (Reykjavik: University of Iceland Press, 2009), 171-195.

[39] Michael T. Taussig, The Devil and Commodity Fetishism in South America (Chapel Hill, NC: University of North Carolina Press, 1980).

[40] John Lindow, "The Male Focus of Scandinavian Household Spirits," in Papers IV: The 8th Congress for the International Society for Folk Narrative Research, ed. Reimund Kvideland and Torunn Selberg (Bergen: Folkekultur, 1975), 35-46.

[41] Lauri Honko, "Memorates and the Study of Folk Belief," Journal of the Folklore Institute 1 (1964): 5-19.

CA Journal of Cultural Analytics

Per's reaction is understandable.[42] With the current labeling, this story is undiscoverable to a researcher interested in field reapportionment, hunting/poaching, household spirits, and the relationships that bind these elements together. Through the use of methods developed for computational folkloristics, this story -- along with many more liminal stories -- becomes more readily discoverable. Given these difficulties in classification, where stories of interest to researchers may be lost in the immensity of the collection, we turn our attention to finding new ways to "surface" the uncertainty of story classification to aid in retrieval.

Probabilistic classification

To get at the problem of classification and to make use of the messy boundaries inherent in folklore classifiers, we construct a probabilistic classifier that simulates Tang Kristensen's decision process: based on reading a story, the task of the classifier is to assign it to a category. Tang Kristensen's two collections, *Danske sagn* (DS) and *Det jyske Almueliv* (JA), include a total of 36 top-level classifications [Table 1], each with multiple second-level categories (657 in DS, and 118 in JA). We represent stories as unordered "bags of words." Note that we distinguish between word *tokens*, which are individual instances of a word in a particular position in a story, and word *types*, which are distinct strings of characters.

Our goal is to find a mathematical formula that represents the association between word types and categories. Due to its simplicity and robustness, we select the naïve Bayes (NB) model. We divide each story into a sequence of word tokens by finding groups of consecutive Unicode (UTF-8) letter characters. We do no further preprocessing of the vocabulary, such as stemming or stopword removal, so very frequent words still appear in the corpus.

After tokenization, we create a word count histogram for each category. We collect all the stories in a given category and record the total count of each word type within those stories. At the end we have 36 word count histograms, one for each category (or 775 for the second level categories). There is evidence that Tang Kristensen carried out this same procedure manually, placing each fair-copied story on one of several piles, which he subsequently subdivided into smaller piles.

We estimate the probability of a word appearing in a category by dividing the number of times the word occurs in the category by the total number of word tokens from stories in that category. For example, if there are 1000 tokens in the category *Household spirits* and the word *nisse* appears 20 times, the word has a 0.02 probability.

Given a story, we can assign a probability score to each category by multiplying the probability of the individual word tokens under that category. We do not vary the probability of words based on their position in the story: the model assumes that each word in a sequence is an independent event that depends only on the category, and not on any other previous word in the sequence. This independence assumption is the *naïve* part of the model. Although the model is clearly not accurate, it is simple and surprisingly powerful. A more practical problem is that individual word probabilities tend to be small numbers, and multiplying small numbers by other small numbers results in extremely small numbers. As a result, rather than multiplying probabilities we add the logarithms of probabilities, which is numerically more stable.

The resulting function represents the relationship between words and categories, but we are also interested in the relationship between individual documents and categories. For example, we might look for documents that are "misplaced" or that could have fit in several categories. We simulate the folklorist placing a new story in one of several piles by removing each story in turn from the data set and evaluating the log probability of each class. Which pile would the classifier select?

The NB classifier makes this "leave-one-out" evaluation easy because the model consists only of word counts. Removing the effect of a document from the classifier is simply a matter of subtracting the token counts from the histogram. We first remove the count for all the words that occur in the document from the original category, evaluate probability scores for each category, and then add the words that occur in the document back into the original category. In most cases, the original classification is the most probable for the left-out story, but we frequently find that another category has higher probability: the story is, at least according to this method, "misclassified".

Indeed, four stories specifically mentioned by Tang Kristiansen himself as being questionably categorized have higher probability in a different class, though not always the one he specified. So, for instance, JFm 4-388, which Tang Kristensen classified as "On prophecy and portents" but felt could just as easily been placed with "fairy tale legends", was classified by our system as a religious legend. JFm 4-389, which Tang Kristensen also classified as "On prophecy and portents" but felt

[42] Abello et al., "Computational Folkloristics."

CA Journal of Cultural Analytics

could have been classified along with stories about ”revenants and all types of ghosts” was classified by our system under ”Hauntings” with the word *skikkelse* [shape] being particularly influential upon the reclassification. Two stories that Tang Kristensen classified as ”religious legends” but felt could just as easily have been classified as witchcraft, JFm 4-438 and 439, were placed by the NB classifier into categories with stories about the ”Devil” and ”Cunning Folk.” In the latter story, the words *kjælling* [hag], *råd* [advice], *dygtig* [clever], *posen* [bag] , and *præst* [minister] were particularly influential.

Once we have estimated the association of each story with each category *as defined by all other stories*, we can reconstruct connections between categories. A common means for describing the behavior of an automated classifier is through a confusion matrix. In this representation, each class corresponds to one row and one column. Rows represent classes as labeled by the annotator, and columns represent classes as predicted by the classifier. The value of the cell at row *i* and column *j* shows the number of stories that have been manually labeled with class *i* but have been placed by the classifier in class *j*. We display this matrix visually using a grid of circles. The area of each circle represents the number of stories in the corresponding cell of the confusion matrix, with blue circles for stories where the human classifier and the algorithmic classifier agree, and red where they disagree. Figure 4 shows a representation of the confusion matrix for the top-level categories. Reassuringly, in all cases the cell on the diagonal is the largest in each row, indicating that the human and the algorithm often agree. We see that classes vary in size: stories about hidden folk are more prevalent than stories about

wyverns.

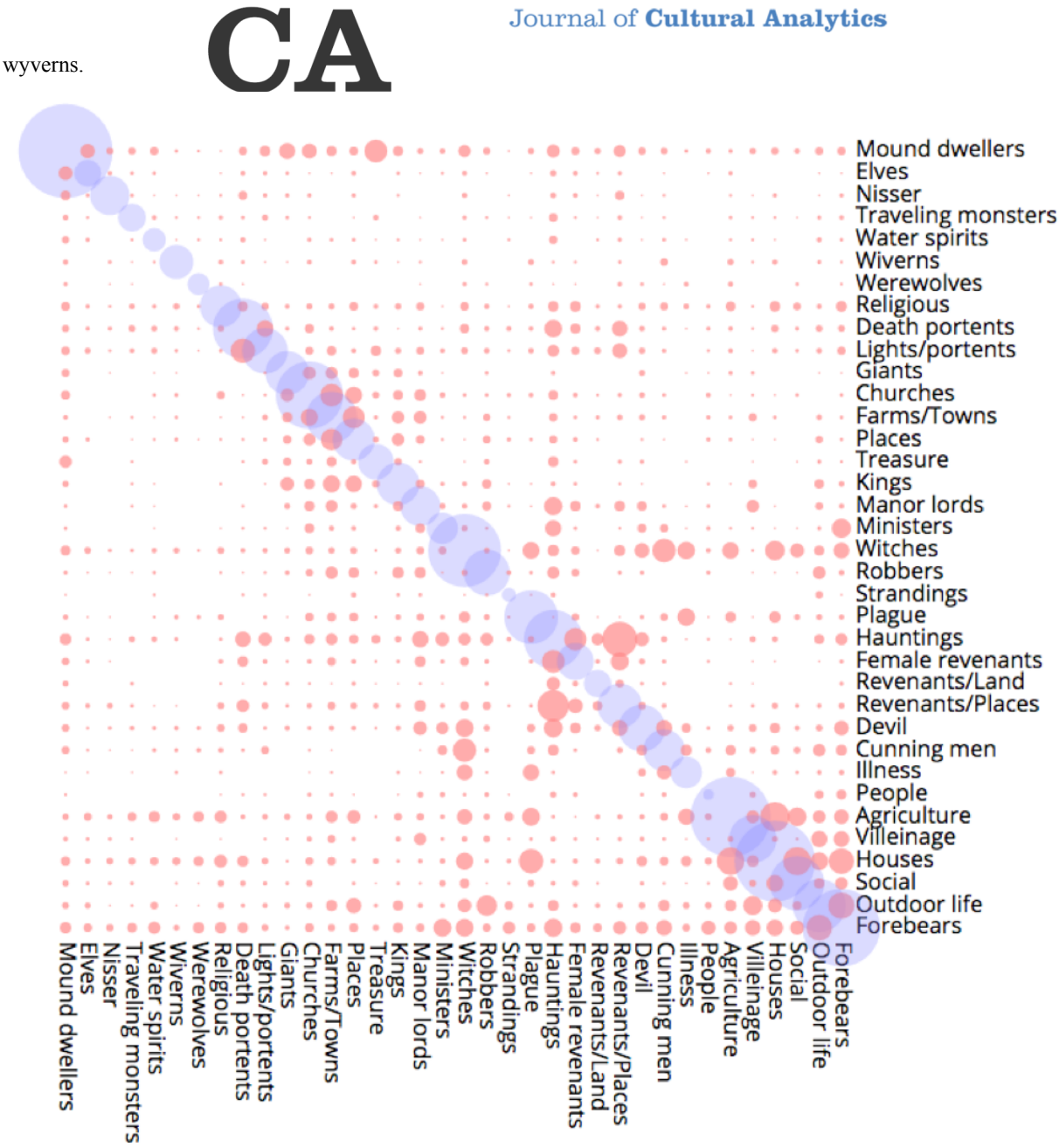

Figure 4. A ”confusion matrix” generated from the 31,088 Danish legends collected by Evald Tang Kristensen, with the size of the circle in each element indicating the number of stories classified into one of 36 high-level categories by Kristensen (rows) and by a NB classifier trained on Kristensen’s classifications (columns). Red circles indicate ”disagreement” between Kristensen and the NB classifier, while the blue circles along the primary diagonal indicate agreement between the training set and the model.

Original label	Stories w/orig. label	Stories w/label from NB	Matches	Precision	Recall	F-score
Mound dwellers	2409	2199	1785	0.8117	0.741	0.7747
Elves	255	294	146	0.4966	0.5725	0.5319
Nisser	428	392	312	0.7959	0.729	0.761
Traveling monsters	257	289	159	0.5502	0.6187	0.5824
Water spirits	199	277	111	0.4007	0.5578	0.4664

CA Journal of Cultural Analytics

Original label	Stories w/orig. label	Stories w/label from NB	Matches	Precision	Recall	F-score
Wiverns	329	337	234	0.6944	0.7112	0.7027
Werewolves	145	197	98	0.4975	0.6759	0.5731
Religious	693	575	353	0.6139	0.5094	0.5568
Death portents	1113	1127	726	0.6442	0.6523	0.6482
Lights/portents	842	676	418	0.6183	0.4964	0.5507
Giants	545	654	385	0.5887	0.7064	0.6422
Churches	1288	1353	916	0.677	0.7112	0.6937
Farms/Towns	891	1155	530	0.4589	0.5948	0.5181
Places	652	878	360	0.41	0.5521	0.4706
Treasure	411	510	256	0.502	0.6229	0.556
Kings	668	684	366	0.5351	0.5479	0.5414
Manor lords	651	759	323	0.4256	0.4962	0.4582
Ministers	450	460	196	0.4261	0.4356	0.4308
Witches	1830	1771	1074	0.6064	0.5869	0.5965
Robbers	680	754	426	0.565	0.6265	0.5942
Strandings	81	156	40	0.2564	0.4938	0.3375
Plague	893	1018	585	0.5747	0.6551	0.6123
Hauntings	1602	1718	698	0.4063	0.4357	0.4205
Female revenants	607	670	274	0.409	0.4514	0.4292
Revenants/Land	260	302	153	0.5066	0.5885	0.5445
Revenants/Places	897	1086	397	0.3656	0.4426	0.4004
Devil	1030	787	444	0.5642	0.4311	0.4888
Cunning men	753	779	348	0.4467	0.4622	0.4543
Illness	424	494	202	0.4089	0.4764	0.4401
People	125	162	22	0.1358	0.176	0.1533
Agriculture	2154	1763	1294	0.734	0.6007	0.6607
Villeinage	706	789	455	0.5767	0.6445	0.6087
Houses	2363	1888	1314	0.696	0.5561	0.6182
Social	889	994	597	0.6006	0.6715	0.6341
Outdoor life	1287	1143	564	0.4934	0.4382	0.4642
Forebears	2279	1996	1211	0.6067	0.5314	0.5666

Table 1. Classification statistics for the Danish legends corpus, comparing the stories assigned by Tang Kristensen to each of the 36 high-level categories to the classifications made by a naïve NB classifier trained on Tang Kristensen's assignments.

Patterns in the off-diagonal elements indicate relationships between classes. Some classes are easier to detect than others. Looking across each row, we can see how easily the algorithmic classifier can find stories identified by the human classifier. The "werewolves" category is fairly clean, probably because it involves specific terms such as "werewolf". In contrast, the "hauntings" category overlaps considerably with several categories about revenants.

Looking at multiple lines at once also reveals clear "block" structures across certain categories. Blocks occur when the original labeling assigns stories to groups of adjacent categories in ways that the NB classifier is unable to distinguish clearly, resulting in greater levels of disagreement ("confusion") among the adjacent labelings, which are visible as clumps of red circles (see Figure 5). These blocks are often situated along the main diagonal, but may appear elsewhere. For example, legends that one might call "ghost stories" (those concerned with hauntings, revenants, etc.) are numerous and poorly distinguished by the existing categories and overlap, not surprisingly, with stories about the devil. This last overlap is, of course, intriguing, as ghosts could only be made theologically "sound" after the Protestant reformation by aligning them with Satan.[43] The most confusing classes for the algorithmic classifier, however, are the final six volumes of *Jyske Almueliv*,

[43] Timothy R. Tangherlini, "'Who ya gonna call?': Ministers and the Mediation of Ghostly Threat in Danish Legend Tradition," Western Folklore 57, no. 2/3 (1998): 153-178.

CA Journal of Cultural Analytics

which form a largely distinct sub-corpus in the overall collection and have a more coarse-grained set of classification labels.

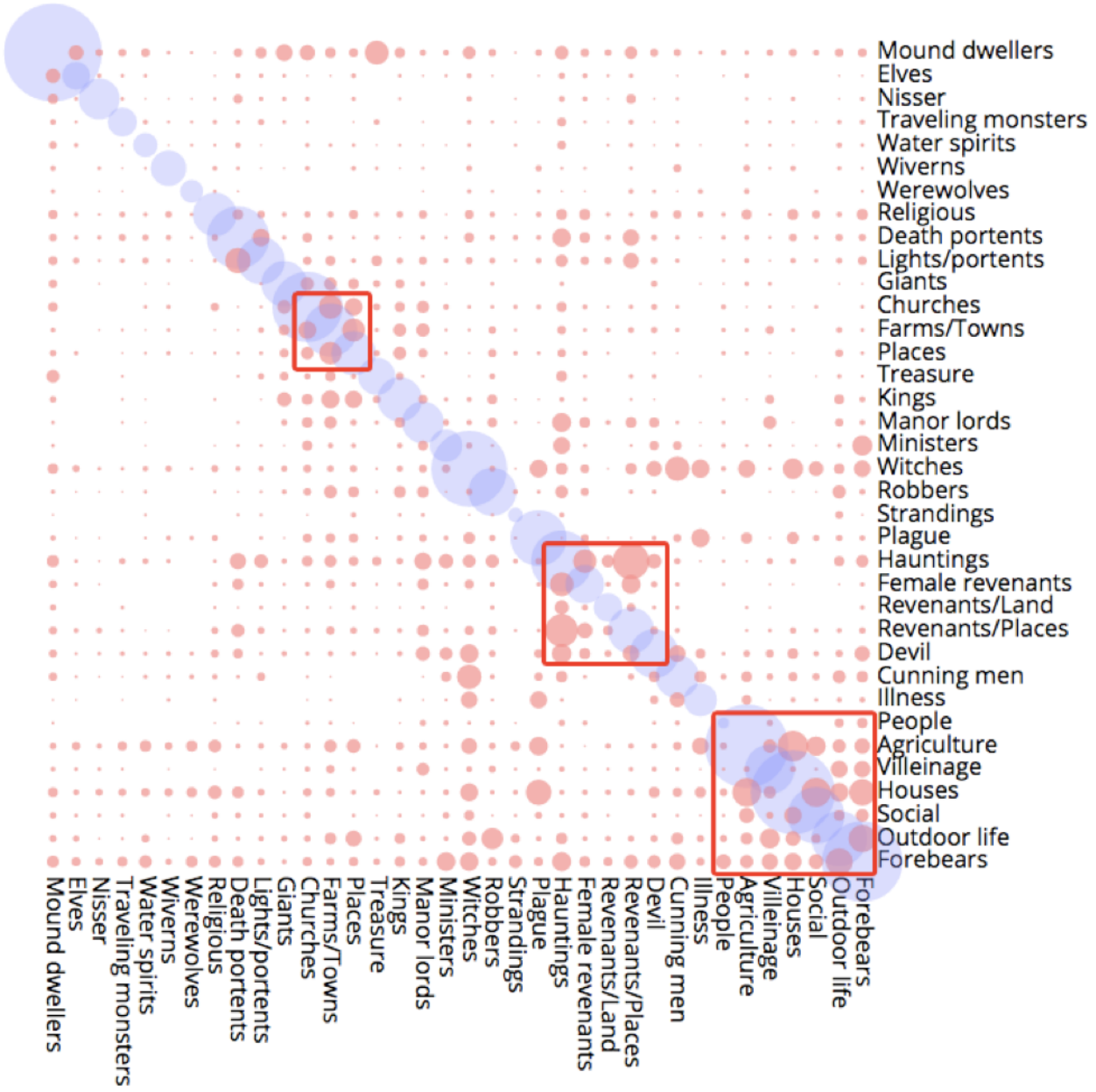

Figure 5. Highlighted ”blocks” in the confusion matrix, indicating recognizable large-scale groupings of stories across adjacent categories. In general, stories within such a grouping will be more similar to each other than they are to stories outside of the block.

Note, however, that ”blocks” of similar categories may not be visible if overlapping categories are positioned far from each other in the matrix. For instance, witches, plague, and illnesses also would produce a noticeable block structure, signaling related content in their stories, if their labels were ordered adjacently. It is also important to look for lone ”hot spots” of confusion between associated phenomena: for example, between hidden folk / mound dwellers and treasure.

In our online interface, users can click on any of the circles to select a pair of classes and view documents where there is disagreement between the human classification and the algorithmic classification. We display both the stories themselves along with the words that are the most significant indicators of one class or another, using a binomial log ratio test,[44]

[44]This test measures the significance of the difference in the frequency of a word between two sets of documents. We compare two hypotheses: first,

CA Journal of Cultural Analytics

specifically Dunning's g-test.[45] This test is sensitive to the number of observations of a given word: a small difference in usage ratio for a very frequent word may be strong statistical evidence, while the same ratio for a word that occurs only a handful of times is indistinguishable from purely random assignment. An example of this interface is shown in Figure 6, for stories classified by Tang Kristensen as "mound dwellers" that were classified by the algorithm as "churches".

Figure 6. A detailed view showing stories that were assigned one classification (Bjærgfolk -- hidden folk / mound dwellers) by Tang Kristensen and a different classification (Churches) by the NB classifier. This view is accessed from the confusion matrix by selecting the point at the intersection of the row corresponding to "Mound dwellers" and the column corresponding to "Churches" (see inset). The color-coded list of words at top highlights the words that the NB classifier found to be indicative of each category, color-coded along a spectrum from red to blue to indicate the intensity of their relevance to either the first category (Hidden folk, in red and orange) or the second category (Churches, in green and blue).

that the probability of the word is the same in both sets and that any observed variability is due only to random chance, and second, that the probability of the word differs across the two sets. If the word occurs x times out of N words in one set and y times out of M words in the second set, we are comparing the probability of these two sets under two binomial distributions with a single proportion p = (x + y) / (N + M) to the probability of each set having its own binomial distribution with proportion p = x/N or p = y/M, respectively.

[45] Ted Dunning, "Accurate Methods for the Statistics of Surprise and Coincidence," Computational Linguistics 19, no. 1 (March, 1993): 61-74.

CA Journal of Cultural Analytics

We can use the same methods to inspect the second-level classes defined by Tang Kristensen. Figure 7 shows a confusion matrix generated from 2,409 stories from the first volume of the collection, concerning mound dwellers or ”hidden folk”. The structure of the overall corpus is mirrored in the structure of this volume. As with the first-level classes, the algorithmic classifier frequently agrees with the human classifier, although there are some classes, such as ”mound dwellers fight,” where the classifiers never agree. We can also see patterns of ”mistakes” in the form of block structures along the diagonal. These indicate that there are distinct sections within the volume, but that Tang Kristensen’s classification scheme is more fine-grained than the bag-of-words lexical representation can support. For example, there are several second-level classes related to ”peel board and rake” and several classes related to the destruction of mounds that show noticeable lexical overlap. Looking at second-level classes in subsequent single volumes in the collection also reveals a similar pattern of block structures, but this pattern becomes less prominent with each volume. A similar pattern emerges in the classifications in *Jyske Almueliv* (see Figure 8).

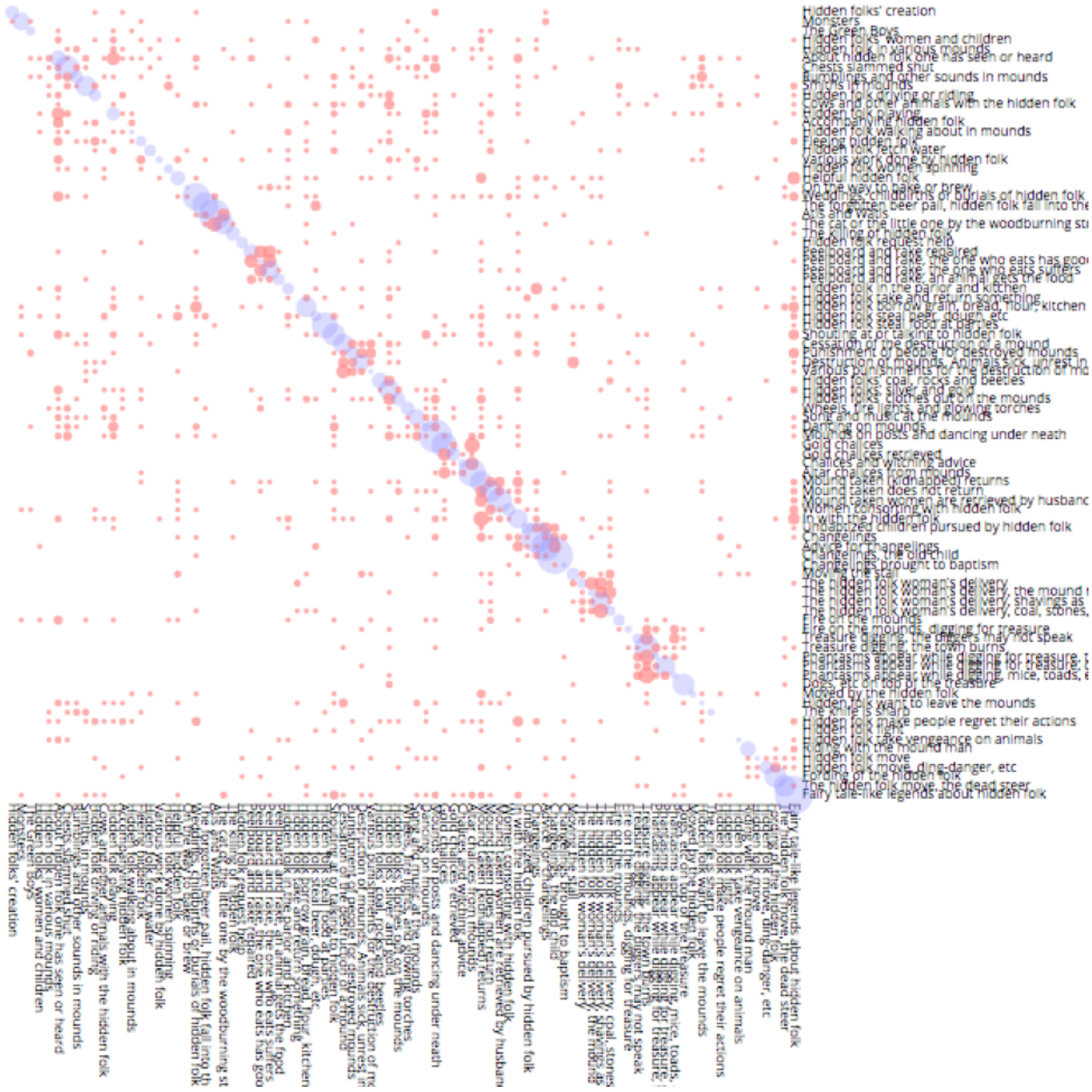

Figure 7. A ”confusion matrix” generated from the 2,409 Danish legends that Tang Kristensen assigned to the category Bjærgfolk (Mound dwellers/hidden folk). The elements of the matrix correspond to the intersections of the stories assigned to one of 86 secondary categories by Tang Kristensen (rows) and by a NB classifier trained on Tang Kristensen’s

CA Journal of Cultural Analytics

classifications (columns).

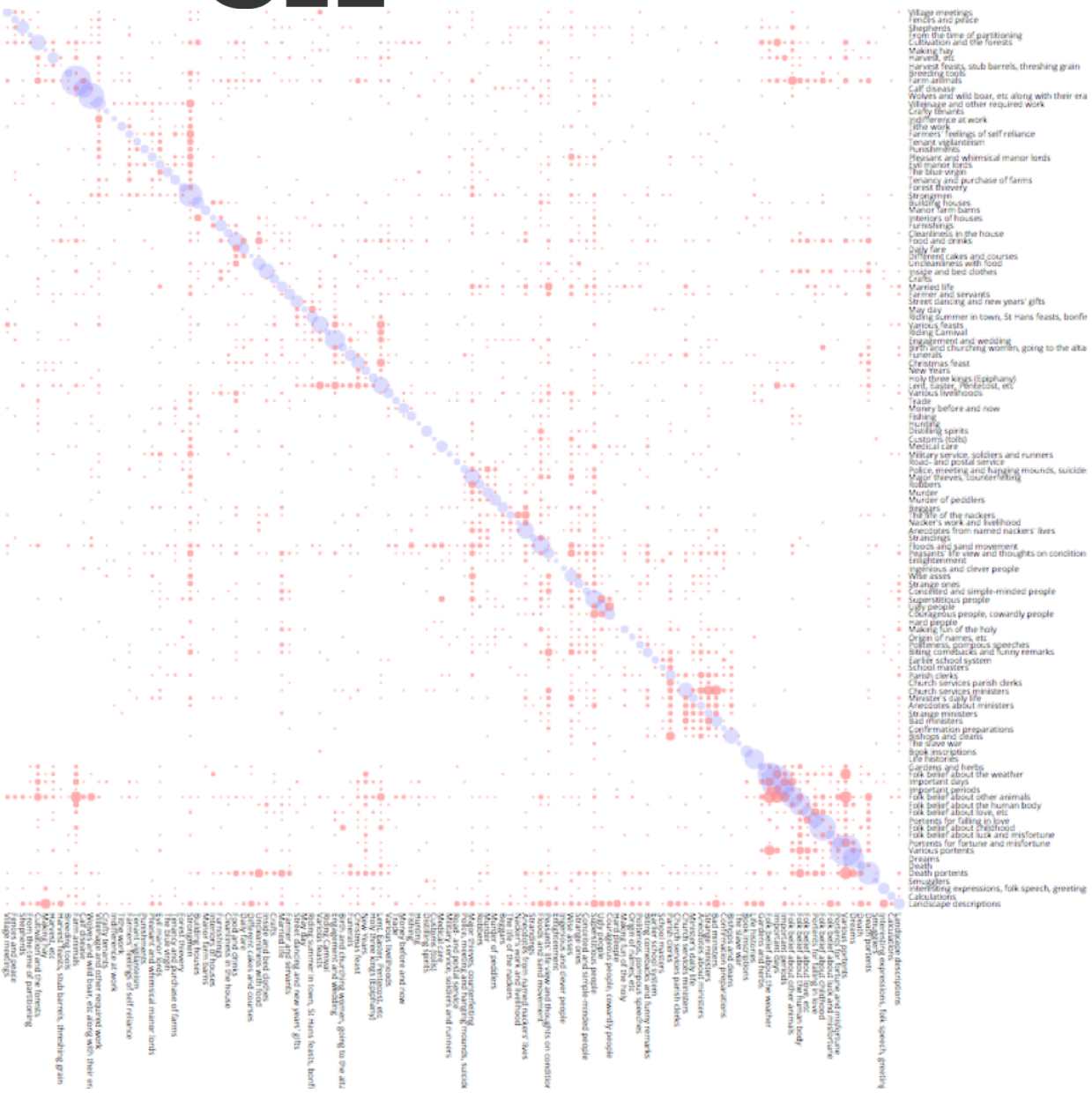

Figure 8. A confusion matrix generated from the 9,679 Danish legends in Tang Kristensen's six-volume collection Det jyske Almueliv (JA). The elements of the matrix correspond to the intersections of the stories assigned to one of the volumes' 118 secondary categories by Tang Kristensen (rows) and by a NB classifier trained on Tang Kristensen's classifications (columns).

Discussion

This confusion-matrix based approach to classification brings to the fore the challenges that Tang Kristensen and Thompson confronted as they attempted to put stories into piles or motifs into poker chip slots, while sidestepping the limiting factors associated with these older one-story/one-label systems. Importantly, it keeps the existing classification accessible, yet allows for exploration across the now-porous boundaries between categories. Several examples of "misclassified"

CA Journal of Cultural Analytics

documents--or documents classified by Tang Kristensen one way, and by the NB classifier another way--illustrate the productiveness of this approach.

A classification method such as the one presented here is only worth pursuing if it allows one to do something otherwise impossible. In this case, finding the liminal stories that shift across existing categories provides two huge advantages. First, a researcher interested in a particular category can rapidly see which other categories might include stories of interest--an impossibility in the standard analog approach to collection navigation. For example, a person interested in hauntings can see that the ”churches” category might include a number of stories of interest. In most cases, modern readers would not know that such affinities might exist. Second, clicking through to the contested classifications allows one to discover stories that occupy these borderlands. Such stories often encapsulate storytellers' shifting conceptions of the relationships between categories that classification schema tend to represent as stable.

In our explorations of the interface, we found numerous stories that revealed the multivalent meanings created through storytelling. The story of the tell-tale hat that opened this essay is but one such example. Stories of manor lords that the probabilistic classifier labeled ”female revenants” are another such productive area of ”misclassification”, and provide numerous examples of the fraught class relationships that structured rural life in 18th and 19th century Denmark. On the one hand, wealthy manor lords exploit far poorer women while, on the other hand, status considerations make relationships based on love essentially impossible:

> They say that at Egeskov manor, a young woman has been entombed in a wall, she was in love with the stall hand, and now they haunt.[46]

These stories--otherwise difficult to discover, if not essentially hidden in the collection--help paint a far more nuanced view of the economic tensions between classes. An exploration of stories about witches classified as cunning folk (as noted above), and vice-versa, provides intriguing material to support investigations of people's conceptualizations of the difference between malevolent witches and helpful cunning folk, revealing that the boundaries between these categories were highly porous.[47]

The disruptions posed by our system to the existing classification schemes also highlight aspects of editorial practice, bringing to the fore the largely etic thinking that governed Tang Kristensen's (and most other) topic classifications. Consequently, the approach can play an important role in interrogating both editorial and scholarly practice not only at the time of initial classification, but also at the time of subsequent re-classification. A simple modification to the underlying collection allows one, for instance, to compare the confusion matrices generated for the first collection of legends, *Danske sagn*, and the second collection of legends, *Danske sagn, ny række*. Do the types of misclassifications hold steady through time, or does Tang Kristensen have shifting ideas about which story to place in which pile?

Our goal with this classification method can be seen as the inverse of usual machine learning classifiers. Unlike in standard machine learning, where one has a high degree of confidence in the training set, in this case we do not. Instead, we are interested in highlighting the uncertainty that Tang Kristensen says he felt at the very moment of classification: ”Should I place this story with those of mound dwellers, or with those of *nisse*?”, one can almost hear him asking. In our approach, stories can straddle this boundary, thus recognizing the polysemy of any story. It also enables a researcher to read any story in the context not only of the stories with which it is classified, but also in the context of other stories with which it could have been classified. This broadening of the context in which stories can be understood restores agency to the researcher, and begins to (re)capture the emic categories of the storytellers themselves.

The approach is not without its faults, and it is important to consider not only how this analysis can go wrong, but also how one would know that it had gone wrong. If classifications are noisy or inconsistent, the output of the confusion matrix might lead one to propose a relationship across categories where none exists. It is relatively easy to diagnose this problem, as the interface allows one to drill-down into any of the category pairs. For example, in the category of stories about ministers classified by the NB classifier as hidden folk, one finds the following story:

> A hired hand who was supposed to dig peat out in a little depression south of the Sørvad mounds got sick of it and stashed his peat spade in a fox hole, ran away, and got to Copenhagen where he started to study. He didn't come back until he was finished with his education and had become a minister. Then he brought the

[46] Tang Kristensen, Danske sagn, vol. 4, 262.

[47] Timothy R. Tangherlini, “‘How do you know she's a witch?’: Witches, Cunning Folk, and Competition in Denmark,” Western Folklore 59, no. 3/4 (2000): 279-303.

CA Journal of Cultural Analytics

> people out to the heath and showed them where he'd hidden the spade. The shaft was completely rotted, but the iron was more or less unharmed.[48]

Here, the mounds where the boy stashes his spade as he makes his escape to Copenhagen to become a minister tricks the automated classifier into thinking that the story concerns hidden folk. Otherwise, these categories have very little to do with one another, yet the confusion matrix indicates there may be something to look at. Similarly, if the pre-existing categories are very similar semantically, something that is a problem with the second-level indices, the NB classifier may misconstrue the attributes that differentiate them. While this does not render the classifier's output useless, as the resulting confusion of its output still functions to highlight the fuzziness of these categories, it does require one to exercise caution. Although the classifier can suggest areas of interesting slippage in the classification, the verification of that slippage still needs to occur manually.

In machine learning, it is common to speak of "ground truth." But, since the ostensible ground truth of Tang Kristensens' classifications is the object of our study, the normal conception of ground truth does not hold. Instead, we may ask, what are we comparing? Potentially, and in keeping with Dundes's ideas of the distinctions between emic and etic categories, the confusion matrix can be seen as charting the fuzziness of informants' own internal classification systems. At the very least, the confusion matrix output captures the polysemous nature of storytelling--stories are never about just one thing. And just as the story about the tell-tale hat discussed earlier in this article was not so much about mound dwellers as it was about the dramatic changes in land use that characterized the late nineteenth century, many of the other stories that straddle categories reveal a similar dialectic tension. It is indeed this productive dialectic between tradition on the one hand and the individual storyteller on the other hand that underlies the folkloric process and animates storytelling. Stories are only told when they are meaningful, and people use stories to explore, negotiate, and solve complex problems in a context of changing and at times nebulous rules. Folklore can be seen as a crowd-sourced representation of the landscape of belief--multiple stories create a terrain that people recognize, while allowing them to negotiate and reconfigure the boundaries of that terrain. It is not surprising, then, that a single story is often hard to classify. In our system, any new story is compared to all the other stories (classification) and fitted into the emerging model of belief. At those places where classification is difficult--the very thing that our system has shown itself adept at identifying--we can recognize that culture is being negotiated.

Conclusion

Classification is a problem that stretches at least as far back as Aristotle. With better, less materially constrained representations of a knowledge domain, we can align classification with the needs of the researcher. Each user can reshuffle even relatively large collections quickly and in a manner that provides information not only about the underlying collection but also about the previous classification regimes. Importantly, materiality no longer limits sorting as it did when these collections were originally classified. In our investigations of the Tang Kristensen collection, we find that the borderlands between categories (where the original annotator picks one category and the probabilistic classifier picks another) are the richest hunting ground for new findings.

Although most scholars do not work with collections of folklore, we also believe that folklore collections are an excellent proving ground for new approaches to classification and indexing that may become increasingly valuable in other fields. Folklore study rests on the underlying idea that there are latent patterns in the variants of traditional expressive forms that constitute the domain. Surfacing these latent patterns can help us understand a great deal about the dynamics not only of storytelling but also of culture. Large, historical collections of expressive forms collected from thousands of storytellers provide an excellent means for identifying the issues that people thought were important and can reveal how storytelling plays a role in the negotiation of culture. Importantly, folklore collections which, by definition, capture popular expressive forms circulating on and across social networks, can be seen as an early analogue to social media, and therefore may be an excellent starting point for studies of cultural production and the circulation of popular ideas such as memes in the current age.

Folklore, given its disciplinary history, highlights the challenges of classification. By questioning existing classification schemata, we can interrogate ideas of classification allowing us to both confirm schemata at the broad level while challenging them at the most detailed level. The methods we have devised for discovering hidden aspects of a collection are

[48] Tang Kristensen, Danske sagn, ny række, vol. 4, 190.

CA Journal of Cultural Analytics

deliberately straightforward (they essentially amount to counting, albeit very very quickly). Indeed, our system can be applied to any corpus that classifies documents into groups. Importantly, we show that this relatively straightforward process can reintroduce someone to a corpus they thought they knew well. As more text, far beyond folklore, becomes available and open to interpretation through computational analysis, more researchers will be in a position to use tools such as the one presented here as a vital means to discover cultural borderlands. We have little doubt that, as people wander through the shifting terrain of expressive culture, they will encounter strange characters, perhaps even some wearing red hats. Now, however, they will be better equipped to answer the question, "I wonder what he is doing here?"